Atypical Event and Typical Pattern Detection within Complex Systems

Brett G. Amidan
Thomas A. Ferryman
Battelle PNWD
902 Battelle Blvd
Richland, WA 99352
509-375-3692
brett.amidan@pnl.gov
tom.ferryman@pnl.gov

Abstract— Algorithms[1] have been developed to find typical patterns and atypical events within complex data systems. A software package called "*The Morning Report*" was developed in which these algorithms were applied to digital flight data for commercial airlines. These systems contain many sets of data with hundreds of variables being measured over time generally resulting in many gigabytes of data to be analyzed. Using statistical and mathematically based algorithms this software identifies atypical flights, along with identifying which flight parameters and which flight phases are atypical. These algorithms also cluster the flights into a finite number of distinct patterns. This allows the flight analyst the opportunity to focus on atypical flights, as well as the typical flight patterns discovered, removing the need to individually explore each flight separately. This software is titled "*The Morning Report*" because it was designed to run each night, producing a report in the morning. This report only identifies the characteristics of the newly added flights, but it uses past flight data to help establish the baseline. The report consists of interactive analysis tools that allow for plotting of significant flight parameters for each atypical flight as compared to the typical flights, as well as plots that contrast a flight pattern of interest to any other flight pattern, or all patterns combined.

The approach, algorithms and software are extendable to a large variety of domains to identify the typical patterns, atypical reports, and providing a plain English explanation.

ACKNOWLEDGEMENTS

Several people made major contributions to this effort. Irv Statler and Tom Chidester of NASA Ames and Loren Rosenthal of Battelle Mountain View envisioned the overall nature of the program and arranged for funding as part of the NASA Aviation Safety Modeling and Monitoring program over the last several years. Paul Whitney and Alan Willse, all of Battelle Pacific Northwest Division, were major contributors. Gary Prothero and his team at ProWorks are also major contributors. They programmed the workstation to process the data and apply the algorithms described in this paper.

TABLE OF CONTENTS

1. INTRODUCTION

One of the main tasks within the Aviation Performance Measurement System (APMS) program is the software development of "*The Morning Report*." *The Morning Report* contains mathematical algorithms that analyze digital flight data. These analyses include statistical methodologies to find atypical flights and establish typical patterns. *The Morning Report* process is divided into 3 distinct phases: (1) Data Transformation, (2) Analysis, and (3) Displays and Plots.

Phase 1 of *The Morning Report* focuses on transforming the flight data for analysis in *Phase 2*, display in *Phase 3* and efficient storage. The number of flight parameters for different aircraft ranges from 70 to more than 400. The data consist of continuous (interval ratio) parameters (such as, airspeed and roll) and categorical (discrete) parameters (such as air ground switch and autopilot mode). Most of the parameters are recorded every second, but some are measured up to eight times a second or once every two seconds. With thousands of flights a day and hundreds of parameters being recorded usually every second for each

[1] Patent pending.

flight, the amount of data escalates into the gigabytes and the ability to analyze the data becomes more difficult.

In order to analyze large amounts of flight data, certain data analysis techniques were employed. Processing was designed to process all the information from the flight in one pass, so that raw flight data will not continually need to be accessed. Analysis was designed to focus around the partitioning of each flight into specific flight phases. These flight phases are then subdivided into subphases. These phases and subphases have been designed to allow for the analysis of a time period during a flight that is relatively homogenous within each flight and nominally similar between flights.

Within each flight subphase, a mathematical signature that represents the data characteristics of the flight is calculated and stored. These signatures serve two purposes– summarize the data characteristics of the flight during the subphase while greatly reducing the storage size, and allow for graphical representations of the data, without having to store and access all the data.

The summarizing signatures store information about each flight parameter concerning such areas as magnitude, rate of change, and data variability for continuous data. Discrete flight parameters are also summarized by characterizing the time spent in each state, and frequency of state transitions (i.e. landing gear going from up to down).

Once flight signatures are calculated they are stored in a static database. These signatures become the inputs for *Phase 2* of *The Morning Report.*

Phase 2 is the analysis phase of *The Morning Report*. It is designed to be run as often as an analysis is desired. It is also designed to run larger analyses overnight, so that they can be viewed in the morning, hence the name "Morning Report".

The first step of *Phase 2* is to select which flights will be included in the analysis. Characteristics that may be used in defining your selection of flights include: departing and/or landing airport and/or runway, date of flight, flight phases (i.e. cruise), flight parameters of interest, and type of equipment. (For over night processing, a standard set of characteristics is typically used.)

The next step is to determine typical patterns. The mathematical signatures are used, along with a statistical clustering algorithm to group similar flights together. The groups are established within each flight phase. Those flights that are not like other flights usually do not cluster into any of the groups and are left as "singletons". The clustering technique currently employed uses K-means clustering [1]; however, there are many others that could have been selected, each having its own benefits and disadvantages.

After similar flights are grouped together, then an algorithm is applied to determine which flights are atypical. An atypical flight is a flight that, for one reason or another, is different than the other flights. This difference may be obvious, such as airspeed of 200 mph, when all of the other flights have airspeed close to 150 mph. The difference may be more subtle by involving multiple parameters that by themselves are not atypical, but are when studied collectively. An example of this is a flight that may have airspeed that is modestly different from the mean value and a vertical speed that is also only modestly different than the mean value; however, it could be atypical if a flight has both at the same time and in a pattern that contrasts with the others.

Atypicality is determined using a series of mathematical computations to determine an atypicality score. The scores are calculated within each flight phase, as to keep the flight segments homogeneous. These atypicality scores are then converted to a scale that is not dependent on the flight phase. This allows for the flight/flight phases to be ranked in order of most atypical to least. The top 1% of atypical flight/flight phases are defined as level 3 atypicalities. Then the next 4% are defined as level 2 atypicalities. The next 15% after the first 5% are defined as level 1 atypicalities. The percentages selected are arbitrarily but reflect the nominal number of atypical flights that will be referred to the aviation expert as atypical.

After these computations are completed, the cluster membership information and atypicality scores are stored into a static database. They become the inputs for *Phase 3*. *Phase 3* allows the user to decide what they want to display and then allows the use of drill-down displays and plots to help the user understand the characteristics of each atypical flight or group of flights, as well as which flight parameters are influencing the atypical flights.

Although *Phase 2* is described as being processed on digital flight data, it is interesting to note that it really could be processed on any type of data. Atypicality scoring and clustering are independent of the source of data. Using other sources of data may require a different mathematically representative signature to input into *Phase 2*. It may also result in different plots in *Phase 3* that are more suited to the type of data.

The following sections will explain the mathematical data signatures that are calculated in the first phase of *The Morning Report*. This will be followed with sections explaining the *Phase 2* data analyses processing, including atypicality calculations, and finding typical groups. A discussion of *The Morning Report Phase 3* displays and plots will then follow.

2. Representative Mathematical Signature

Because most parameters are recorded every second and flights can last hours, a large amount of data needs to be sifted through. One method to help with this is dividing each flight into flight phases. Each of these flight phases is then subdivided into subphases. These phases and subphases allow the investigator to compare many flights during similar portions of flight. This may show that a flight was atypical at take-off, but typical during the other phases.

These phases and subphases have been designed to allow for the analysis of a time period of data during a flight that is more homogenous. Ten phases have been used in analyzing flight data. These phases are titled: Taxi Out, Takeoff, Low Speed Climb, High Speed Climb, Cruise, High Speed Descent, Low Speed Descent, Final Approach, Landing, and Taxi In. Each phase can consist of multiple subphases. For example, the Landing phase consists of the subphases – initial touchdown to reverse thrusters on, reverse thrusters on to 80 knots, and 80 knots to 60 knots.

The objective of the *Phase 1* processing is to summarize each parameter of each flight for each phase or subphase with a mathematical signature. This processing summarizes the many seconds of data for a parameter that occur in a given phase for each flight, providing only a few data statistics that summarize the data characteristics mentioned earlier. These signatures are calculated for both continuous data parameters and discrete parameters. The continuous data signatures will be discussed first.

The first step in creating a continuous data mathematical signature for a given flight and parameter is using 5 seconds of data before and after the data at the first second of the desired phase to create a vector of 11 seconds of data. Then fit a quadratic least squares model, $y = a + bt + ct^2 + \varepsilon$, with time as t, the vector of data as y, and the coefficients of a (intercept), b (slope), and c (quadratic). (To facilitate interpretation, we used a centered version of this model.) An ε (error) is calculated as a vector of difference between the actual y values and the predicted y values. The ε vector is converted to a single value, $d = [(\Sigma\varepsilon^2)/(\text{n-3})]^{1/2}$. Repeat this process with the 2nd second of the same phase and continue to the last second of the same phase. Each second will now have a corresponding set of coefficients (a, b, c, and d). Each of the four statistics has a value for each time record in the phase. If the phase is 10,000 seconds long (as cruise might be), 10,000 sets of a, b, c and d's are calculated. Then each coefficient is summarized by calculating the mean value, standard deviation, minimum value, and maximum value. This calculation results in a mathematical signature for a flight in which the phase data of each parameter is summarized into 16 statistics (4 coefficients with 4 statistics each). Then the parameters value at the beginning of the phase and at the end of the phase is added to the signature, resulting in an 18 statistics summary of the parameter during a particular flight phase. Appendix A contains a technical discussion of the steps involved in the continuous data signature.

To help readers conceptualize this mathematical signature, the following comments are presented. Hopefully, they help to clarify and give examples of what some of the statistics mean. The statistics calculated for the coefficient a represent the magnitude of the parameter values. If a flight has a higher mean a value than the other flights for the parameter airspeed, then its airspeed is on the average higher than the other flights for that given phase. The statistics calculated for the coefficient b represent the rate of change (slope) of the parameter values. If a flight has a higher mean b value for the parameter airspeed than the other flights, this means its airspeed is increasing at a higher rate than the other flights for that given phase. The statistics calculated for the coefficient c represent the curvature or rate of rate of change of the parameter values. If a flight has a higher mean c value for the parameter airspeed than the other flights, then its airspeed change in the slope is occurring at a higher rate than the other flights for that given phase. The statistics calculated for the statistic d represent the amount of variability in the parameter values. If a flight has a higher mean d value for the parameter airspeed than the other flights, then that means its airspeed is more variable than the other flights for that given phase. Its airspeed is fluctuating high and low more than the other flights. Other words that are conceptually equivalent to the coefficients a, b, c, and d are *average value*, *velocity*, *acceleration*, and *noise*.

The phase processing for the continuous parameters results in a signature matrix that has n rows and 18*p columns, where n is the number of flights (1 row for each flight) and p is the number of parameters (18 statistics for each parameter).

Additional columns are added to the signature matrix which contains the discrete parameter signatures. Discrete parameters contain different modes that are categories with no reasonable mathematical calculations possible between the varying modes. Example discrete parameters include: thrust reversers, landing gear, air ground switch, slats, etc.

The signature for discrete parameters is calculated from a transition matrix. A transition matrix shows how many times a mode of a parameter changes to another value for the next record and how long the data for a given flight phase remains in each mode. The transition matrix consists counts of time periods during the flight phase. It is converted to a related matrix form by dividing the off-diagonal counts by the total number of counts for the flight phase; result in the off-diagonal cells reporting the percentage of time in which that flight was in that mode during a particular flight phase. The diagonal of the matrix reports actual counts that the flight changed from one mode of the parameter to another, found in the off-diagonal elements of the matrix. The transition matrix is a square matrix with dimensions equal to *n* (the number of possible modes). After a transition matrix is created it needs to be "vectorized". This means that an *n* x *n* matrix will become a vector with n^2 elements. This vector is the signature for this categorical parameter and is added to the signature matrix.

The continuous and discrete signatures are used in *Phase 2* for analyses. A different type of signature is also needed for graphical representation of the flights during *Phase 3*. This signature allows for plotting the trend of a parameter over time. This signature is referred to as the data compression signature, because it is similar to the raw data; however the size has been greatly reduced. We use one compression technique for categorical data, run-length encoding, and a different technique for continuous data, progressive linear interpolation (PLI).

The run-length encoding is a standard data compression technique that stores the value and time of the data variable at the time of any change. It is loss-less compression.

For continuous variables, the data compression signatures provide a smoothed fit through the raw data without missing the unusual values. This reduces the number of data points stored from thousands (the raw data) to only a few hundred (the data compression data). This allows for smaller data storage, and quicker access for plotting.

The PLI method begins by using only the first and last data points, and connects them with a line. The residuals are then calculated and the (x,y) point associated with the largest absolute residual is then selected as an intermediate point. Line segments are then formed from the first data point to the intermediate point and from the intermediate point to the last data point. Residuals are recalculated and the data point associated with the largest new absolute residual is selected as an additional intermediate point. The process is repeated until a pre-specified accuracy level is attained or until the number of data points used to establish the linear segments reaches some pre-specified limit.

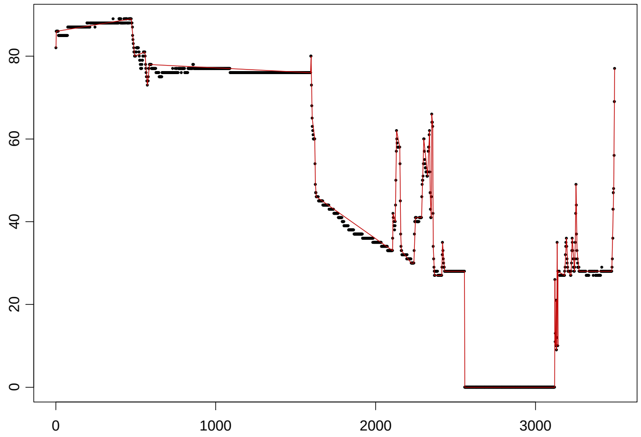


Figure 1 – Example of the fit obtained with the PLI method.

Figure 1 shows an example of how well the PLI method estimates the raw data for a given parameter. In this example there were over 3000 raw data points. The PLI method found the optimal 100 data points that when the points are connected, it fits the raw data with minimal residual. The data compression signature gives a good estimate of the magnitude of the flight parameter over time. One weakness of the data compression signature is when the flight parameter has a lot of variability and/or bounces around a lot. The data compression signature may not fully display all the variability.

An important aspect of the signature process is that the intent is for the primary analysis to be done from the characterization signature process and the compression signatures will be used for display. It is possible, with modest loss of accuracy, to do analysis based on the compression signatures. It is recommended to the users to keep the raw data for the possible, but hopefully very rare, event that detailed analysis is necessary and not supportable from the characterization signature or the compressed data signatures.

The signature process can be I/O and CPU intensive but is is easy to set it up to run continuously through out the day and if necessary on multiple machines. Our experience has been that the processing requirements are easily met with a single pc.

Once the signatures are calculated for an individual flight they are stored in a database. The discrete and continuous signatures are then available for the analysis phase of *The Morning Report*, while the data compression signatures are available for the plots and displays in *Phase 3* of *The Morning Report*. The raw data from the flight is no longer needed in the subsequent steps of *The Morning Report.*

3. CLUSTERING METHODOLOGY

Phase 2 is the analysis phase of *The Morning Report*. The analyses performed include clustering and atypicality determination. It is designed to be run as often as an analysis is desired. It is also designed to run larger analyses overnight, so that they can be viewed in the morning, hence the name "Morning Report".

The first step of *Phase 2* is to select which flights will be included in the analysis. Characteristics that may be used in defining your selection of flights include: departing and/or landing airport and/or runway, date of flight, flight phases (i.e. cruise), flight parameters of interest, and type of equipment.

The next step of *Phase 2* is clustering the mathematical signature data. The goal of the cluster analysis is to identify clusters, or groups, of both typical and atypical flight patterns. The ideal cluster analysis should 1) identify flight patterns, 2) enable the characterization of those patterns, and 3) provide insight into the typicalness or atypicalness of each cluster.

There can be thousands of flights daily, making it impossible for a flight expert to study each flight individually. When similar flights are grouped together into a smaller number of groups, then the expert can focus on the characteristics of the group. Each pattern or group will have a set of common characteristics in which the expert may be able to classify that group with a label such as "unstable approach". This allows the expert to focus on specific groups of flights, and not having to investigate every individual flight.

Statistical clustering techniques are applied to the discrete and continuous signature matrices calculated in *Phase 1*. Although Kmeans clustering is used in *The Morning Report*, many clustering techniques exist and could be applied in grouping together similar flights. In general, the clustering technique assigns each flight into the group in which its signature is similar to the other members of the group. When a flight's signature is unlike the signatures in each of the groups, then it is assigned to its own group.

If a flight is the only member in a group, then it is referred to as a "singleton". A flight may be a singleton because it is mathematically between established groups and it isn't close enough to be considered a member of any group. These flights maybe called "inliers" if they lie between the clouds of data, but not actually within a cloud.

Another possible way a flight can be a singleton is that it has extremely low or high signature data values. These flights are called "outliers". They lie on the outside of all the clouds of data.

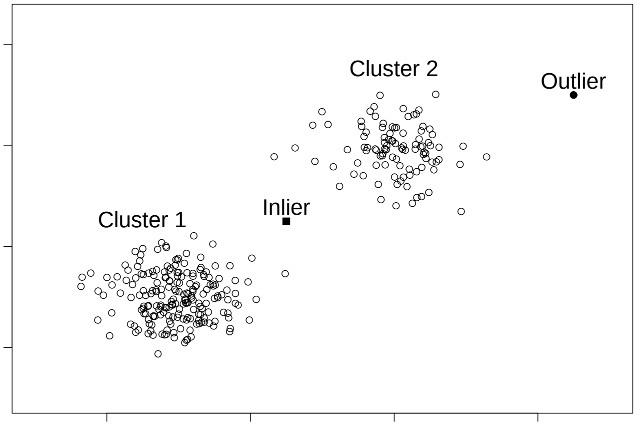


Figure 2 – Example of clusters and singletons

Figure 2 shows a two dimensional plot of the concept of an inlier and outlier. The inlier lies between clusters 1 and 2 and is different than the two groups. It may look different than the two groups; however it lies close to the average of all the data and may not be too alarming. The outlier lies beyond cluster 2. It is different than the two groups and its value is quite different than the average of all the data. In the next section we discuss atypical flights. The outlier in this example would be considered the most atypical because it is both a singleton and it is furthest from the mean.

4. ATYPICALITY METHODOLOGY

It is common practice for aviation experts to find and investigate safety issues with flights. Without some help in identifying which flights to investigate, they will look at hundreds of normal flights for every flight they find that is intriguing from a safety aspect. This process can be time-consuming. Furthermore, they must rely on their own expertise and perceptiveness. They must compare the data to their own standards of acceptable performance to determine if a flight displays a problem. *The Morning Report* uses statistical and mathematical methods to present the expert with a set of flights that are atypical and consequently much more likely to have noteworthy flight data to study. It also focuses the attention of the expert on the flight phase and parameter(s) that are most likely to offer insight to aviation safety issues. In turn, the tool will identify problems that expert may have envisioned as possible problems and problems that the expert may never have envisioned. This can help the expert to *think outside the box* and to find more obscure but important insights that have been hidden in the data and thus facilitate improved aviation safety.

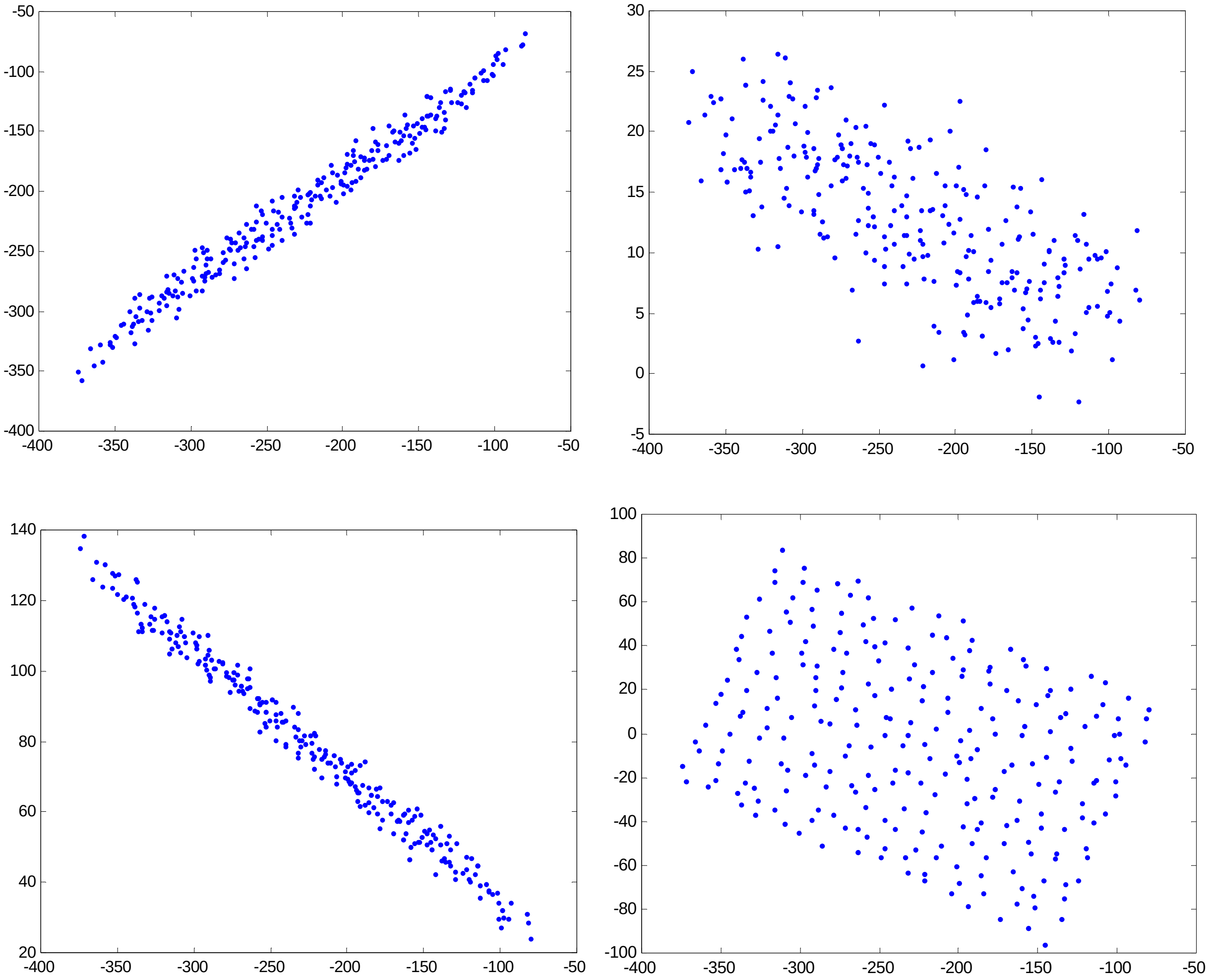


Figure 3 - Plots Showing Various Data Presentations of the Multivariate Test Data

Atypicality calculations are performed on the same flights that were selected at the beginning of the *Phase 2* processing. Atypicality calculations are made on the same mathematical signatures calculated in *Phase 1*. They are done by flight phase. This section will explore the concepts of the atypicality methodology. A technical explanation of this method can be found in Appendix B.

One of the main statistical procedures used in atypicality determination is principal component analysis (PCA) [2]. This method transforms the variables into eigenvectors. The method is useful in combining the information found in many variables into just a few transformed variables. An example of this process is found in Figure 3. The data for this example consists of 256 records with 16 variables each. The user might choose to investigate the nature of the data by plotting the records using variable pairs as the axes. These four plots show four pairs of these variables being plotted against each other. The relationship between the data points cannot be seen clearly in any of these plots. Figure 4 shows a plot of the records using the first two principal components from the PCA of the original data. This plot shows a clear pattern that could not be found in the previous plots because of the multivariate nature of the data. In this case, PCA is able to make order out of what appears to be chaos.

PCA takes the original parameters and transforms them into linear combinations. This reduces the dimensions of the data from many parameters (hundreds or even thousands) to only a few transformed parameters called PCA components (usually under 100, depending on the correlation structures in the original data). This reduction is accomplished by finding eigenvalues and corresponding eigenvectors for the data, then ordering the eigenvalues from largest to smallest. Each eigenvector is then multiplied by the original data to create a PCA component. The number of PCA components is the same as that of the variables; however, the first components contain much more information than the last components. Only the components in which the eigenvalues (starting from largest to smallest) sum to 90% of the total of all the eigenvalues are kept. These eigenvectors represent 90% of the variance observed in the data. The other data are thought to represent random variability in the data and no significant loss of information occurs when they are discarded.

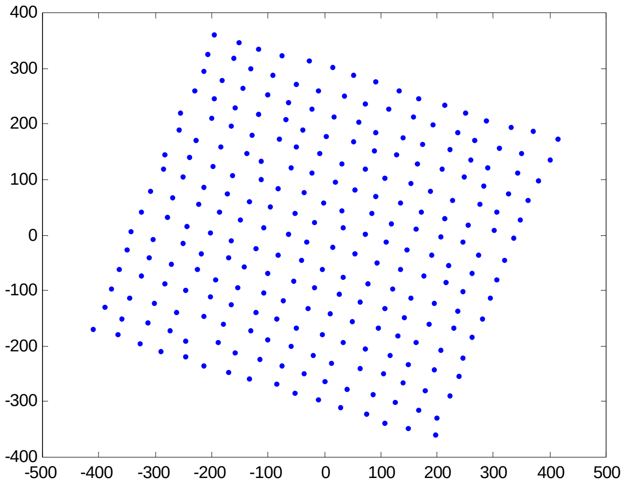


Figure 4 - Plot Showing the Nature of the Figure 1 Data

The PCA components could now be plotted against each other. These plots may show outliers in the data, or possible grouping among the data. Figure 5 shows an example of a 2-dimensional scatterplot of two PCA components. Data points can be seen that are far from the pack of data, indicating those flights are atypical from the others.

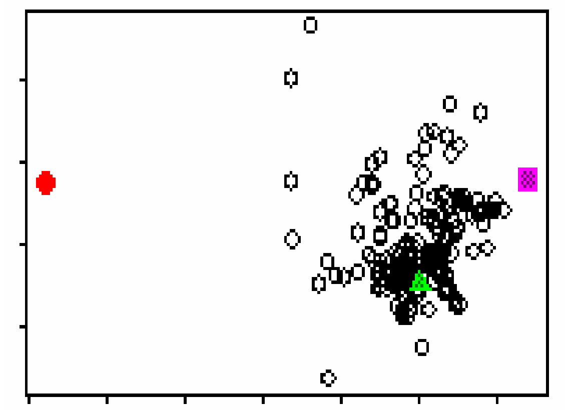


Figure 5 - A Scatterplot of Two PCA Components

From the PCA components, an atypicality score is calculated using Mahalanobis distance [3]. The atypicality score is calculated for each flight using the following formula:

$$A_i = \sum_{j=1}^{n} PCA(j)_i^2 \; / \; \lambda_j \qquad (1)$$

where i is the flight (row) in which the atypicality score is calculated; A is the atypicality score for a flight; $PCA(j)_i$ is the jth PCA component vector and ith element in the vector for $j = 1$ (corresponding to largest eigenvalue) to n (corresponding to the smallest eigenvalue not cut-off by the 90% rule); and λ_j is the eigenvalue associated with the PCA component of interest. An atypicality score close to zero indicates a typical flight, and a large atypicality score, relative to the other flights, indicates an atypical flight for that specific flight phase.

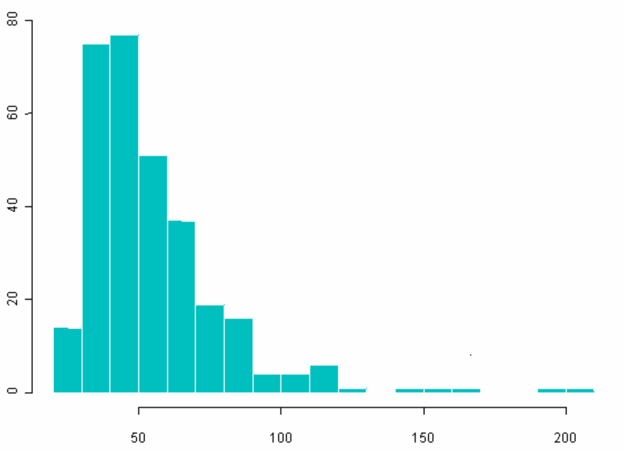


Figure 6 – Example histogram of atypicality scores

A histogram of the atypicality scores displays the value for the flights and allows easy identification of atypical flights. Figure 6 shows a histogram of the atypicality scores. In this example, the flight with the largest score (furthest right) is most atypical for that particular flight phase and should be the first flight investigated. The atypical flights shown in the histogram will generally correspond to the atypical flights seen when plotting each of the PCA components (Figure 5).

Atypicality scores are calculated for each flight phase. Most analyses involve multiple flight phases; however the scores are not comparable when looking across flight phases. A method was created to standard atypical scores so that they atypicality could be ranked across many flight phases. This resulted in the calculation of a global atypicality score.

The global atypicality scores are calculated using p-values from the atypicality scores and cluster membership information. Experience has shown that atypicality score histograms (as in Figure 6) usually have a skewed shape with a long tail to the right. It was found that a gamma distribution fit the atypicality scores well, especially in the tail. Although it didn't fit as well with the more typical scores, the purpose of the atypicality scores was to concentrate on the tail. Therefore, a gamma distribution was used to calculate a p-value for each of the atypicality scores.

As was discussed in section 3, cluster membership plays a role in atypicality. A flight that is a singleton is considered to be more atypical than a flight that belongs in a typical cluster. In order to allow cluster membership to play a role in the atypicality process, a cluster membership score is calculated by using the following equation:

$$cms_i = \frac{n_i}{N} \tag{2}$$

where cms_i is the cluster membership score for flight/flight phase i, n_i is the number of flights in flight i's cluster, and N is the total number of flights in the analysis.

The global atypicality score is then calculated using the following equation:

$$G_i = -\log(p_i) - \log(cms_i) \tag{3}$$

where G_i is the global atypicality score for flight/flight phase i, p_i is the p-value for flight/flight phase i, and cms_i is the cluster member score, as shown in equation 2. This results in a global atypicality score that is always positive, with larger scores meaning more atypical. Global atypicality scores usually range between 0 and 25.

Each of the flights is then ranked according to its largest global atypicality score (across all the flight phases) from highest to lowest score. The flights are then classified into one of four levels. The top 1% of flights are classified as level 3 atypicality, the next 4% of flights are called level 2 atypicality, and the next 15% of flights are called level 1 atypicality. The other flights are considered typical.

After the atypicality scores and clustering are completed, the cluster membership information and atypicality scores are stored into a static database. They become the inputs for the *Phase 3* displays and plots.

5. MORNING REPORT DISPLAYS AND PLOTS

Phase 3 allows the user to decide which *Phase 2* analysis they would like to view. *The Morning Report* software then presents tables and plots to help the user understand the *Phase 2* results.

Table 1 is a view within *The Morning Report* that shows the atypical flights in order of atypicality. For each atypical flight, the table lists which flight phase the atypicality occurred and it displays a list of the parameters that most contribute to its atypicality. A paragraph is written by the software explaining how these parameters are contributing to the atypical nature of the flight. The paragraph is displayed elsewhere in the software outputs. These parameters are determined by using simple univariate calculations based on the signature matrix. It then allows the user to drill-down a series of displays and plots to see how the flight parameters are influencing the atypical flights.

These plots consist of traces of the particular flight parameter over time for the particular flight phase, with a backdrop consisting of a performance envelope of other flights. The performance envelope is a contour plot that consists of superimposed gray to black boxes displaying the number of flights that shared that value at that time. More flights in a box are represented by a darker color. Figure 7 shows an example of a performance envelope, as well as traces of two atypical flights. This plot does not represent real flight data; it is for illustrative purposes only. From the plot, Flight 314 looks atypical for flap position because its values are consistently lower than the other flights. It also appears to fluctuate greatly and may be considered atypical for that reason too. Flight 278 also looks atypical.

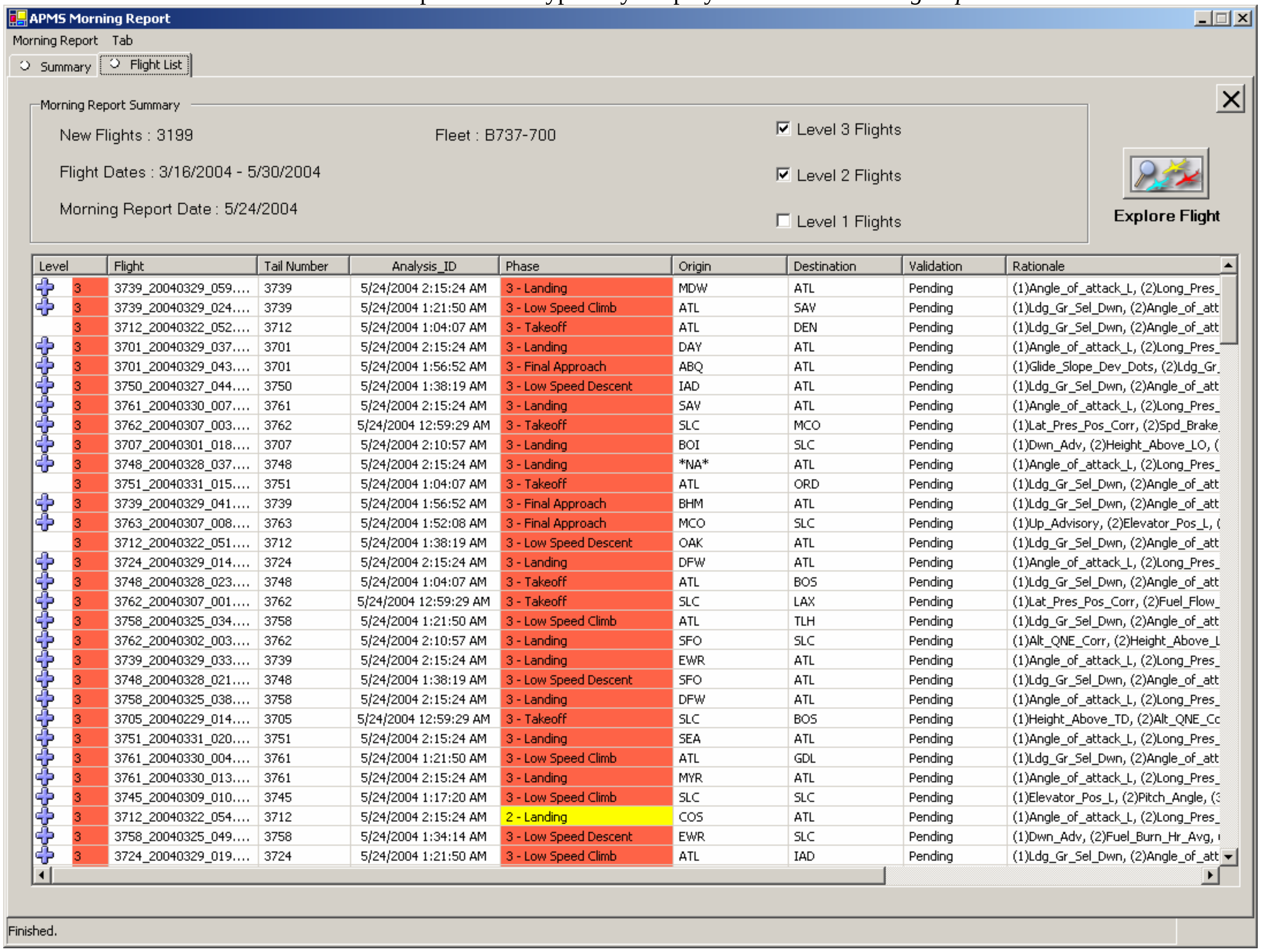
Performance envelopes are also used to contrast two clusters. Figures 8 and 9 show performance envelopes contrasting a cluster of interest and a reference cluster. The reference cluster can be the largest cluster or some other cluster, or it would be the most typical 80% of flights, or even all flights. Figure 8 shows one of the smaller clusters, referred to as an atypical cluster. It shows that the N2 Right values for the atypical cluster are much higher than the reference cluster values during approach. Figure 9 shows that the cluster of interest (typical cluster) has a smaller pitch angle than the reference cluster at takeoff, but the clusters pitch angles become similar when the landing gear is up.

6. CONCLUSIONS

A three phase program has been produced that finds atypical data and typical patterns within a complex system. Although this paper describes applying these methods to digital flight data, it has been extended to other data systems, including air traffic control data, and text report data. It can continue being adapted to new systems. The three phases run independently from each other, allowing for parallel processing. It also makes it conducive to changes when applying it to new data systems.

The Morning Report has proven capable of identifying possible atypical flights that are atypical due to characteristics in the data. This does not remove the domain expert for the equation. This is a tool that can focus the expert's attention and allow for the expert to drill down to the core of the atypicality. It also allows for greater understanding of the general data patterns that exist in the system.

Table 1. Example of the Atypicality Display within *The Morning Report*



Morning Report Summary

New Flights : 3199

Fleet : B737-700

Flight Dates : 3/16/2004 - 5/30/2004

Morning Report Date : 5/24/2004

☑ Level 3 Flights

☑ Level 2 Flights

☐ Level 1 Flights

Level	Flight	Tail Number	Analysis_ID	Phase	Origin	Destination	Validation	Rationale
+ 3	3739_20040329_059....	3739	5/24/2004 2:15:24 AM	3 - Landing	MDW	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3739_20040329_024....	3739	5/24/2004 1:21:50 AM	3 - Low Speed Climb	ATL	SAV	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
3	3712_20040322_052....	3712	5/24/2004 1:04:07 AM	3 - Takeoff	ATL	DEN	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3701_20040329_037....	3701	5/24/2004 2:15:24 AM	3 - Landing	DAY	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3701_20040329_043....	3701	5/24/2004 1:56:52 AM	3 - Final Approach	ABQ	ATL	Pending	(1)Glide_Slope_Dev_Dots, (2)Ldg_Gr_
+ 3	3750_20040327_044....	3750	5/24/2004 1:38:19 AM	3 - Low Speed Descent	IAD	ATL	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3761_20040330_007....	3761	5/24/2004 2:15:24 AM	3 - Landing	SAV	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3762_20040307_003....	3762	5/24/2004 12:59:29 AM	3 - Takeoff	SLC	MCO	Pending	(1)Lat_Pres_Pos_Corr, (2)Spd_Brake_
+ 3	3707_20040301_018....	3707	5/24/2004 2:10:57 AM	3 - Landing	BOI	SLC	Pending	(1)Dwn_Adv, (2)Height_Above_LO, (
+ 3	3748_20040328_037....	3748	5/24/2004 2:15:24 AM	3 - Landing	*NA*	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
3	3751_20040331_015....	3751	5/24/2004 1:04:07 AM	3 - Takeoff	ATL	ORD	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3739_20040329_041....	3739	5/24/2004 1:56:52 AM	3 - Final Approach	BHM	ATL	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3763_20040307_008....	3763	5/24/2004 1:52:08 AM	3 - Final Approach	MCO	SLC	Pending	(1)Up_Advisory, (2)Elevator_Pos_L, (
3	3712_20040322_051....	3712	5/24/2004 1:38:19 AM	3 - Low Speed Descent	OAK	ATL	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3724_20040329_014....	3724	5/24/2004 2:15:24 AM	3 - Landing	DFW	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3748_20040328_023....	3748	5/24/2004 1:04:07 AM	3 - Takeoff	ATL	BOS	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3762_20040307_001....	3762	5/24/2004 12:59:29 AM	3 - Takeoff	SLC	LAX	Pending	(1)Lat_Pres_Pos_Corr, (2)Fuel_Flow_
+ 3	3758_20040325_034....	3758	5/24/2004 1:21:50 AM	3 - Low Speed Climb	ATL	TLH	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3762_20040302_003....	3762	5/24/2004 2:10:57 AM	3 - Landing	SFO	SLC	Pending	(1)Alt_QNE_Corr, (2)Height_Above_L
+ 3	3739_20040329_033....	3739	5/24/2004 2:15:24 AM	3 - Landing	EWR	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3748_20040328_021....	3748	5/24/2004 1:38:19 AM	3 - Low Speed Descent	SFO	ATL	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3758_20040325_038....	3758	5/24/2004 2:15:24 AM	3 - Landing	DFW	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3705_20040229_014....	3705	5/24/2004 12:59:29 AM	3 - Takeoff	SLC	BOS	Pending	(1)Height_Above_TD, (2)Alt_QNE_Cc
+ 3	3751_20040331_020....	3751	5/24/2004 2:15:24 AM	3 - Landing	SEA	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3761_20040330_004....	3761	5/24/2004 1:21:50 AM	3 - Low Speed Climb	ATL	GDL	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att
+ 3	3761_20040330_013....	3761	5/24/2004 2:15:24 AM	3 - Landing	MYR	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3745_20040309_010....	3745	5/24/2004 1:17:20 AM	3 - Low Speed Climb	SLC	SLC	Pending	(1)Elevator_Pos_L, (2)Pitch_Angle, (3
+ 3	3712_20040322_054....	3712	5/24/2004 2:15:24 AM	2 - Landing	COS	ATL	Pending	(1)Angle_of_attack_L, (2)Long_Pres_
+ 3	3758_20040325_049....	3758	5/24/2004 1:34:14 AM	3 - Low Speed Descent	EWR	SLC	Pending	(1)Dwn_Adv, (2)Fuel_Burn_Hr_Avg, (
+ 3	3724_20040329_019....	3724	5/24/2004 1:21:50 AM	3 - Low Speed Climb	ATL	IAD	Pending	(1)Ldg_Gr_Sel_Dwn, (2)Angle_of_att

Finished.

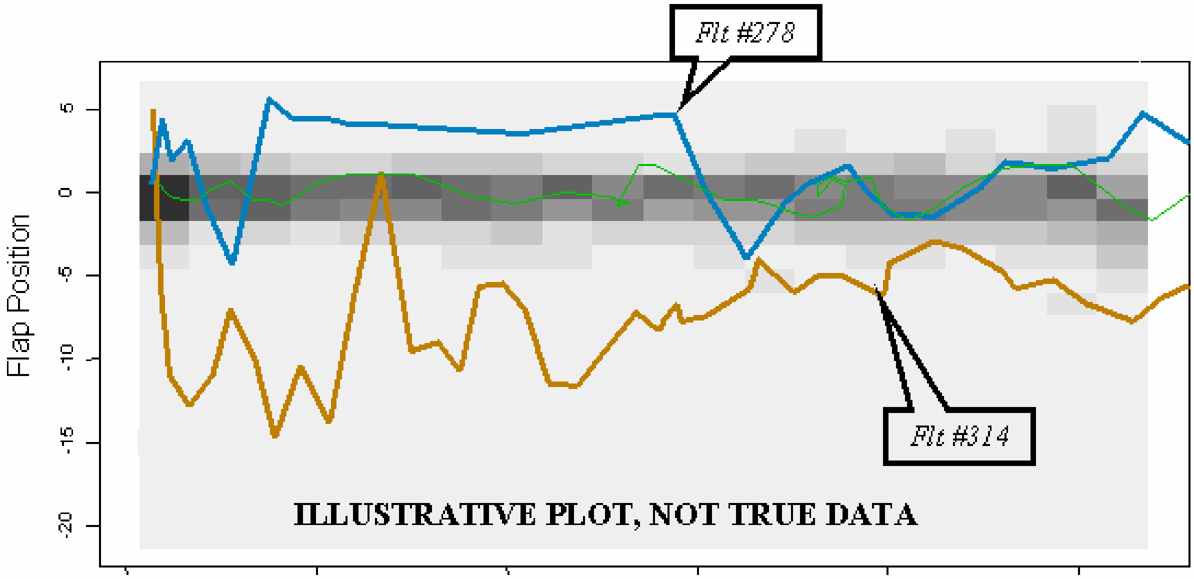


Figure 7 – Performance Envelope Plot of Flap Position with 3 Flight Traces

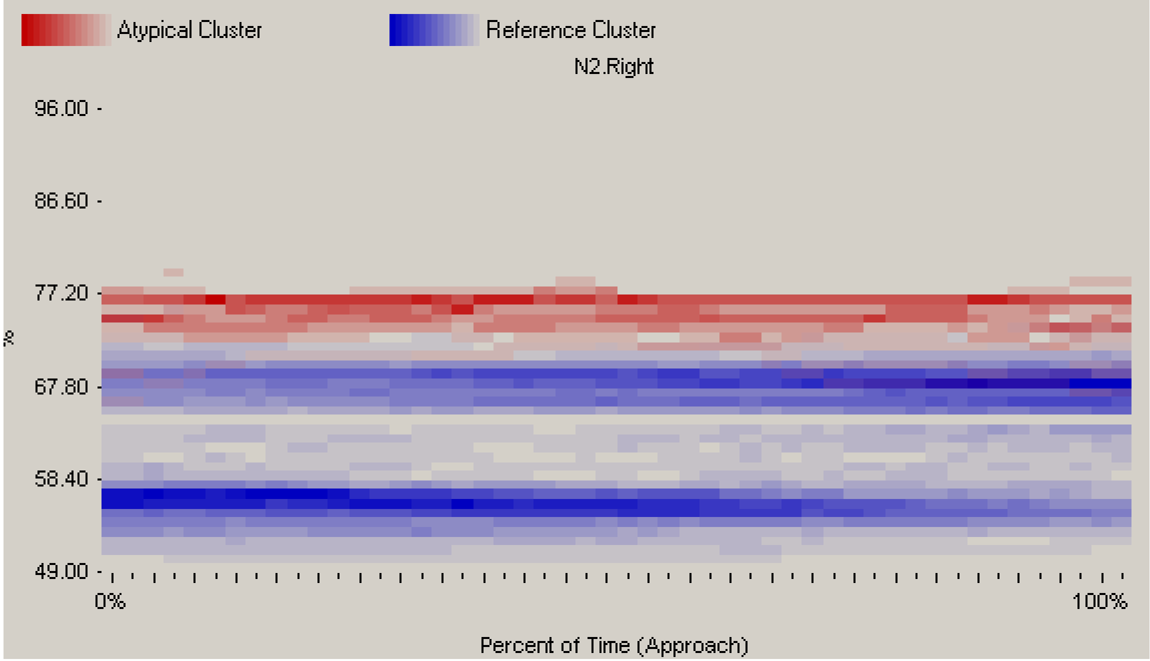


Figure 8 – Performance Envelope with an Atypical Cluster Contrasted with a Reference Cluster

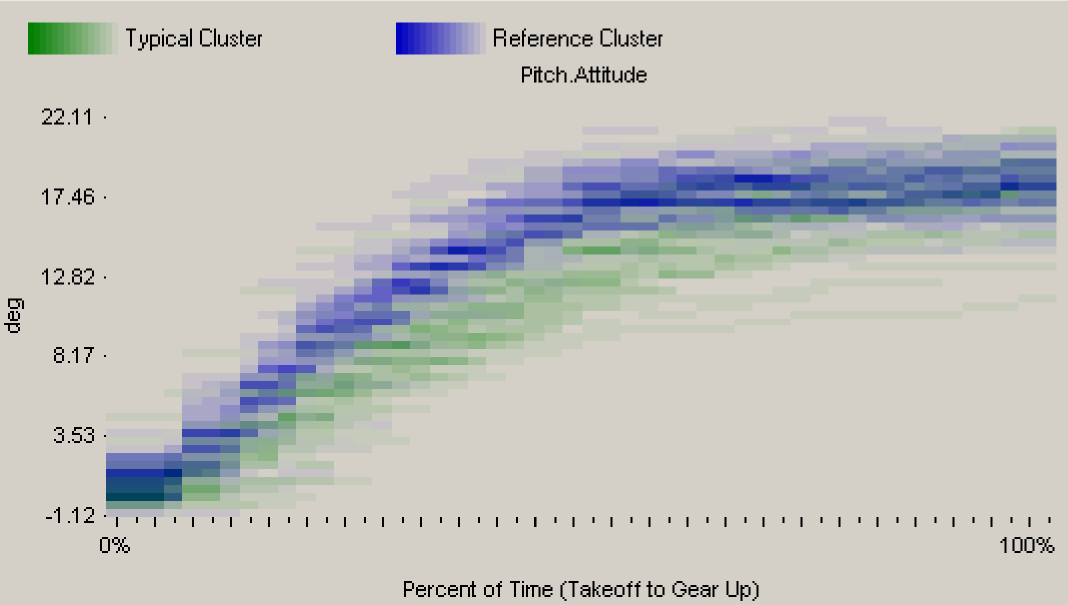


Figure 9 – Performance Envelope with a Typical Cluster Contrasted with a Reference Cluster

REFERENCES

[1] Cluster reference.

[2] Alvin C. Rencher, Multivariate Statistical Inference and Applications, New York: John Wiley and Sons, Inc., 1998.

[3] Geoffrey J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, New York: John Wiley and Sons, Inc., 1992.

BIOGRAPHIES

Brett Amidan *is a Senior Research Scientist at Pacific Northwest National Laboratories operated by Battelle. He has been working in the Aviation Performance Measurement System (APMS) program for NASA Ames for over six years. He has developed and led development of mathematical algorithms applied within the APMS program. He also performs multivariate analysis, experimental design, and simulation efforts on a variety of other projects at PNNL. He previously served as Statistics Coordinator at Clinical Research Associates where his main focus was in experimental design. He has an M.S. degree in Statistics from Brigham Young University*.

Dr. Thomas Ferryman *is a Battelle Chief Scientist at Pacific Northwest National Laboratory with over 30 years of experience in system engineering and mathematics/statistics. He leads the technical development of aviation safety data analysis tools for NASA (numeric, categorical and/or text data). He has also developed prognostic tools for use on gas turbine engines. Prior to coming to Battelle, Dr. Ferryman was Chief Systems Engineer for Lockheed leading a major weapon system modification (AC-130H Gunship).*

APPENDIX A: TECHNICAL EXPLANATION OF THE MATHEMATICAL SIGNATURE CALCULATIONS

Following are the steps used in calculating the continuous mathematical signature done in *Phase 1* processing:

1. Extract the data for a given flight, given parameter, and given phase. Include the data from 5 seconds before and after the given phase. For example, put data into a vector from Flight #1, the parameter airspeed during the touchdown phase (including 5 seconds of data before and after the phase). If *m* is the number of seconds in the given phase, then the extracted data should have *m*+10 seconds.

2. From this extracted data create an 11-second data window with the first 11 seconds (the 5 seconds before the start of the phase, and the first 6 seconds of the phase).

3. Solve the least squares equation: $y = a + bx + cx^2$, with *y* as the data from the 11-second window, *x* as the vector (-5,-4,-3,-2,-1,0,1,2,3,4,5), and *a*, *b*, *c* as the coefficients to be solved. The coefficient *d* is calculated using the equation: $d = \sqrt{\frac{SSE}{(n-3)}}$

   where *SSE* is the sums of squares error (from the analysis of variance table), and *n* is the size of the window (11 in this case).

4. From the data in Step 1, create another 11-second data window by shifting the window over by 1 second, so that seconds 2 through 12 are taken. Then repeat Step 3 with this data and record the coefficients *a*, *b*, *c*, and *d*. Continue to take 11-second windows by shifting over 1 second each time, until the last 11 seconds are sampled. This results in a vector of size *m* for each of the coefficients.

5. Summarize each coefficient vector by calculating the mean, standard deviation, minimum value, and maximum value. This process results in a 16-element vector that summarizes the data for a given flight, given parameter during a given phase. This vector has the following look: $\bar{a}$, s_a, $a_{\min}$, $a_{\max}$, $\bar{b}$, s_b, $b_{\min}$, $b_{\max}$, $\bar{c}$, s_c, $c_{\min}$, $c_{\max}$, $\bar{d}$, s_d, $d_{\min}$, and $d_{\max}$.

6. This process is repeated for each desired flight and each desired parameter and results are put into a matrix with a row for each flight and 16 columns for

each parameter (containing the results from each 16-element vector). This matrix is for a given phase.

7. This resulting matrix can be analyzed to find atypical flights within the given phase using the method explained in Appendix A.

APPENDIX B: TECHNICAL EXPLANATION OF THE ATYPICALITY CALCULATIONS

Following are the steps used in calculating the atypicality scores:

1. Scale the data ($D_{n \times p}$).

$$D_{c(n \times p)} = \left(\frac{d_{ij} - \overline{d}_{.j}}{\sigma_j} \right)$$

where d_{ij} is the element in the D matrix from row *i* and column *j*, $d_{.j}$ is the average of the *j*th column from matrix D, and σ_j is the standard deviation from the *j*th column from matrix D.

2. Remove all rows of data from D_c with at least 1 missing value.

$D_{1((n\text{-}m) \times p)}$ is the matrix with no missing values;
$M_{(m \times p)}$ is the matrix with missing values.

3. Calculate the covariance matrix.

$$C_{(p \times p)} = \text{cov}(D_1) = \left(D_1' D_1\right)^{-1} \sigma^2$$

4. Calculate the eigenvectors and eigenvalues using principle component analysis.

$[E_{(p \times p)}, F_{(p \times 1)}] = Eigen(C)$ where E is a matrix of eigenvectors and F is a vector of the eigenvalues.

5. Truncate the number of eigenvectors to use to *q*, where *q* is the minimal number which satisfies -

$$\frac{F_1 + F_2 + ... + F_q + F_{q+1}}{\sum F} > threshold \quad .$$

After a threshold is decided upon (0.90 was used in these analyses), then *q* number of eigenvectors and values will be kept.

6. Create the new data. Note: D_c is used to create the data, with zeros substituted for the missing values.

$$G_{c(nxq)} = D_{c(nxp)} \times E_{(pxq)}$$

7. Calculate the Atypicality Scores (Mahalanobis Distance).

$$A_i = \frac{1}{q - m'_i} \sum_{j=1}^{q} \frac{G_{c(j)_i}^{\;2}}{F_j}$$

where *i* goes from 1 to n, and m'_i is the number of missing values for that row of data.

8. Calculate the Cluster Membership Scores.

$$cms_i = \frac{n_i}{N}$$

where n_i is the number of flights in flight *i*'s cluster, and *N* is the total number of flights in the analysis.

9. Calculate the Global Atypicality Scores.

$$G_i = -\log(p_i) - \log(cms_i)$$

where p_i is the p-value for flight/flight phase *i*, and cms_i is the cluster member score.